Chapter 4

Rendaku love and hatred: opaque morphological structure[1]

Abstract

Some roots in Japanese compounds always undergo the rule of rendaku, others never undergo the rule, and still others vacillate. In this chapter, we investigate this kind of lexical variation from the perspective of the frequency of such roots. Different types of frequency are considered, such as that of roots in isolation (=token frequency), and the roots' frequency of occurrence as a left- or right-hand member of compounds (=family size and family frequency). We show that frequency matters for the status of roots in the rendaku process. The data in which these frequency effects occur, however, are relatively infrequent, and words with higher frequency undergo rendaku according to the phonological grammar. Very infrequent words thus seem to be exceptions to the rule. Since rendaku clearly involves phonological factors as well as frequency effects, we argue that this result should be interpreted in a model which integrates usage-based factors into phonology.

[1] A version of this chapter has been published as: van de Weijer et al (2013).

Generative and usage-based perspectives sometimes have to be combined in order to account for a particular linguistic pattern (see §1.3). This chapter discusses the Japanese rule of *rendaku* (lit. "sequential voicing") from this point of view. Rendaku voices the first segment of a compound if a number of conditions are satisfied. The history of this rule, as well as its (ir)regularity and the conditions on its application in present-day Japanese, have attracted a great deal of attention in the morphological and phonological literature. One aspect has been especially problematic for past analyses: even if all conditions on its application are taken into account, rendaku appears to have a considerable number of lexical exceptions. That is, some lexical items are more prone to undergo the rule than others, on a seemingly idiosyncratic basis. We will show that the propensity of these roots for rendaku can be related to the frequency with which these roots occur in isolation as well as in compounds. Apart from contributing to the solution of a long-standing puzzle in Japanese morphophonology, a general conclusion that can be drawn from this analysis is that the frequency with which words are used (both in compounds and in isolation), should be included in the explanation for their phonological behaviour. The frequency effect we observe in rendaku neither belongs to Type I frequency effects (analogy), nor Type II frequency effects (reduction), but should be regarded as a Type III frequency effect (opaque structure). Since rendaku is also governed by purely phonological constraints (such as Lyman's Law, see below), a satisfactory solution must combine aspects of theoretical linguistics and of usage-based approaches.

This chapter is organized as follows. in §4.1, we illustrate the basic mechanisms of rendaku and point out some of the relevant conditions on the application of the rule that have been uncovered so far in the literature. In §4.2, we introduce the lexical variation which forms the main topic of this chapter and formulate a number of hypotheses to explain this behaviour. Section 4.3 describes the methodology and the data set we investigated. Section 4.4 provides the statistical results of our investigation. Section 4.5 provides an analysis of the data in EPOT and §4.6 concludes.

4.1 Rendaku

The rule of sequential voicing in Japanese, normally referred to as *rendaku*, has been a topic of discussion in the Japanese and general literature for a very long time (see e.g. Haraguchi (2001), Irwin (2005), Irwin (2009), Itô & Mester (2003), Kubozono (2005), McCawley (1968), Otsu (1980), Rosen (2001), Vance (1987), Vance (2005), Vance (2007), Itô & Mester (1986) and references cited there, also to the extensive literature in Japanese). The basics of the rule are straightforward: the initial voiceless obstruent of a root becomes voiced when the root appears as the right-hand member of a compound (subject to a number of conditions, see below). Examples, taken from various standard sources, are given in (1):

(1) *Examples of rendaku*

k→g	shima	-	kuni	→	shima-guni
	'island'		'country'		'island country'
h→b	roten	-	huro	→	roten-buro
	'outdoor'		'bath'		'outdoor bath'
s→z	maki	-	sushi	→	maki-zushi
	'roll'		'sushi'		'rolled sushi'
t→d	isi	-	tooroo	→	isi-dooroo
	stone		lantern		'stone lantern'

Historically, there was a particle *no* (possessive) between the two compounds, which was reduced and caused the voicing. There are two things to note: the voiced counterpart of /s/ is [dz] (conventionally transcribed as z), and the voiced counterpart of /h/ is [b] (the source for this latter alternation is historical: /h/ derives historically from /p/) (see Frellesvig (2010: Ch.2), Vance (1987: Ch.10), many of the contributions to van de Weijer et al. (2005) and references cited there, for discussion of the history of this phenomenon and other relevant information).

There are a number of conditions on the application of rendaku, some of which are regular and well-known, and some of which are the object of debate. First, rendaku is almost exceptionlessly blocked if there is a voiced obstruent in the second part of the compound itself. This condition on the application of rendaku, which is illustrated in (2), is usually referred to as Lyman's Law (Lyman (1894) and see Vance (1987: 136ff.) and many other sources for discussion).

(2) *Lyman's Law: Rendaku is blocked if it would create two voiced obstruents in one root.*

kami	-	ka**z**e	→	kami-ka**z**e	'kamikaze'
'divine'		'wind'		(kami-*gaze)	

There is also an exception to this generalization, viz. the single lexical item in (3), in which the [g] in the second member does not block voicing of [h] to [b].

(3) *Rendaku is, unexpectedly, not blocked* (Vance 1987: 137).

nawa	-	hasi**g**o	→	nawa-basi**g**o	'rope ladder'
'rope'		'ladder'			

Second, rendaku applies almost exclusively to native Japanese words, and less often to words borrowed from Chinese (which are referred to as Sino-Japanese) or other languages (i.e. older or more recent loanwords). In fact, Martin (1952) claimed that the process is frequent only in native Japanese words, but research since then has shown that there is a considerable number of examples involving Sino-Japanese words, older loanwords (e.g. from Portuguese), and even

some more recent loanwords that undergo rendaku (see Takayama (2005), Irwin (2011: 151f)). This suggests that rendaku is a productive rule. See Vance (1987: 140-141) for examples and discussion, also of the extent to which rendaku applies in mimetics (reduplicated vocabulary) and in other parts of the morphology such as derivation and inflection. In this chapter we will focus on the status of rendaku in native Japanese compounding.

Apart from Lyman's Law and the restriction to loanwords, other conditions on the application of rendaku that have been discussed are related to prosodic size (see e.g. Irwin (2009), vowel length (Horton & Minami 2011), syntactic branching in longer compounds (Otsu (1980), Kubozono (2005)), semantic inclusion/exclusion relationships (Shibatani 1990: 174), and the location of accent (see e.g. Zamma (2005), Ohta (2013), Yamaguchi & Tanaka (2013)). Some of these conditions are (almost) exceptionless (such as Lyman's Law), and others represent tendencies that may block rendaku in a certain percentage of cases, i.e. they are variable. It is also possible that other categories influence the application of rendaku, such as the initial consonant or the initial mora of the root that undergoes the rule: some consonants or moras might be more susceptible for rendaku than others, e.g. because an alternation between [h] and [b] might be less transparent than an alternation between [t] and [d]. Second, the number of syllables of the left-hand side of the compound might matter, or whether the left-hand side of the compound contains a voiced consonant or not (cf. the "extended Lyman's Law", see e.g. Vance (2007). There is isolated discussion of these topics in the literature cited above.

In our investigation, we will focus on disyllabic words where rendaku would be expected, i.e. words in which Lyman's Law is not relevant. We will only take into consideration compounds consisting of two roots, to abstract away from phrasal effects. Similarly, we exclude words with long vowels (Horton & Minami 2011), to avoid any interference of vowel length. Finally, we will not consider semantic aspects of compounds that do or do not undergo rendaku. The properties that were taken into consideration will be spelled out in detail below (§4.3).

First, however, we need to discuss the fact that, even if all of the constraints on rendaku outlined above are taken into account, some lexical items do and other lexical items do not undergo the rule, in a seemingly unpredictable fashion.

4.2 Lexical variation

Individual roots behave differently with respect to rendaku in Japanese. That is, some roots that would at first glance be expected to undergo rendaku do not undergo the process at all, or they show rendaku in some compounds but not in others. Vance (1987: 146) refers to this situation as the "fundamental irregularity" of rendaku, and Miller (1967: 195) describes the (non-)application of the rule as "completely bewildering" (see also Ohno (2000), Rosen (2001)). Vance (1987: 147) gives a number of words that, unexpectedly, resist rendaku, which are reproduced in (4):

(4) *Exceptions to rendaku "never"*

a.	soko	-	tuti	sokotuti	(*soko-duti)
	bottom		soil	'subsoil'	
b.	kutu	-	himo	kutuhimo	(*kutu-bimo)
	shoe		lace	'shoelace'	
c.	suna	-	kemuri	sunakemuri	(*suna-gemuri)
	sand		cloud	'sand cloud'	
d.	yubi	-	saki	yubisaki	(*yubi-zaki)
	finger		tip	'fingertip'	
e.	asa	-	sio	asasio	(*asa-zio)
	morning		tide	'morning tide'	

Note that conditions like Lyman's Law play no role in these forms, since the right-hand members have no voiced segments and all involve native Japanese roots. Thus, they are phonologically fully comparable to the examples in (1). In other words, the fact that they form exceptions to rendaku cannot be attributed to any of the known conditions on rendaku. The question is therefore how these forms should be dealt with: have we missed a condition that is relevant? Are they exceptions that must somehow be marked in the lexicon? Or is the exceptional status of these words due to factors outside the grammar? If rendaku is a productive rule in Japanese, and if no conditions can be found that explain the "exceptions" in (4), then such compounds must be marked as lexical exceptions.

The question whether rendaku is a productive rule or not in contemporary Japanese is relevant here. The issue is discussed in detail by Kubozono (2005). After a consideration of the evidence, partly on the basis of experiments with Specific Language Impairment-speakers (Fukuda & Fukuda 2000), he concludes that rendaku voicing is a productive rule in Japanese, although some cases of voicing (and, presumably, cases where voicing does not take place) may be lexicalized (see also Itô & Mester (2003: 124)). Hence, the question how the non-application of rendaku in forms such as those in (3) should be accounted for is an important one.

Rosen (2001: Appendix D) lists 19 further examples of roots like those in (3) which he characterizes as "never" undergoing rendaku, besides roots which always undergo rendaku ("rendaku-lovers"). He also distinguishes one other category: roots that vacillate between the two options ("rendaku-haters") (citing p.c. from Haruo Kubozono who used similar terms). He also draws attention to the lexical nature of the distinction, i.e. whether a word is a "lover" or a "hater" is unpredictable from any of the (known) conditions on rendaku.

Rosen provides 113 examples of "rendaku-lovers", i.e. words that always or almost always undergo the rule. The relatively large number of "lovers" also suggests that rendaku should be considered a productive rule in Japanese. Examples of rendaku "lovers" are given in (5):

(5) *"Rendaku lovers"*

huro	'bath'	e.g.	uti-buro	'inside bath'
			soto-buro	'outside bath'
			mizu-buro	'water-bath'
kiwa	'brink'	e.g.	yama-giwa	'mountain brink'
			hae-giwa	'receding hairline'
			te-giwa	'skill' [hand brink]
sake	'sake'	e.g.	nama-zake	'raw sake'
			taru-zake	'barrel sake'
			sio-zake	'salt salmon'

(from Rosen (2001: Appendix E))

Both Vance and Rosen note that a number of words vary in their status, i.e. they show rendaku in some compounds but not in others. Rosen refers to such roots as "haters", but we propose to refer to them as "doubters". Vance provides the example in (6a), and Rosen (2001: Appendix F) lists examples of compounds with four of such "doubter" roots, such as those in (6b):

(6) *"Rendaku doubters"*

a. ki	'tree'	cf.	niwa - ki	'garden tree'
		vs.	yama - gi	'mountain tree'
b. hara	'field'	cf.	sino - hara	'bamboo field'
			sasa - hara	'bamboo grass field'
		vs.	una - bara	'ocean field'
			kuwa - bara	'mulberry field'
kusa	'grass'	cf.	ira - kusa	'nettle' [thorn grass]
			natu - kusa	'summer grass'
		vs.	hituzi - gusa	'sheep grass'
			no - gusa	'wild grass'

At first sight, it seems impossible to predict in which category a root might fall, i.e. whether a given compound shows rendaku or not (recall the quotes from Vance and Miller above). However, we suspect that if the notion of frequency is taken into account, it is possible to predict whether roots belong in the category of "never" (4), "lovers" (5), or "doubters" (6), see also our considerations below (8).

We need to define exactly what types of frequency might be relevant in this respect. First, however, it is necessary to take into consideration that in rendaku, a root typically has two allomorphs, one of which is the 'basic' (non-rendaku) form and the other is the rendaku form. These terms are illustrated in (7) (cf. the examples in (5) above):

(7) root: /huro/ 'bath'

allomorphs: [furo] *non-rendaku form*

(i) in isolation
(ii) as left-hand member in compounds

[buro] *rendaku form*

(i) as right-hand member in rendaku compounds

Below, we will investigate whether the frequency of occurrence of both allomorphs of a root is responsible for the variation in rendaku. Consider a root which appears extremely frequently in its *non*-rendaku form (i.e. in isolation or on the left side in compounds) but almost never in its rendaku form. The non-rendaku allomorph will be more strongly represented in the mental lexicon ("entrenched" in Exemplar Theoretical terms) and it is thus possible that it is more likely to impose its shape on the rendaku allomorph (the effect might be compared to a "majority rule" or "paradigm uniformity" effect (see e.g. Kenstowicz (2005)). This would make such roots "rendaku-haters". If roots appear in their rendaku form relatively frequently, we may expect the rendaku allomorph to be sufficiently well-entrenched in order to surface when the phonological conditions for rendaku are satisfied. This would make the root a "rendaku lover". The rendaku forms and non-rendaku forms may also surface in a "reasonable" frequency proportion (where what is "reasonable" should of course be determined on the basis of a statistical investigation). This would make such roots "rendaku doubters".

What types of frequency are relevant here? We will now discuss the probable influence of the frequency of the root in isolation, family frequency of the root as a left-hand member of the compound, family size of the root as a left-hand member of the compound, family frequency of the root as a right-hand member of the compound, family size of the root as a right-hand member of the compound. First, the frequency of a root in isolation can be measured as the number of tokens in a given corpus, and can be determined rather straightforwardly. The number of compounds with a particular root as its left- or right-hand member could be counted in two ways, viz. either as *family frequency* or as *family size*. *Family frequency* refers to the summed token frequencies of compounds in which a particular root appears (regardless of how many different compounds the root appears in). *Family size*, on the other hand, refers to the *number* of compounds formed with a particular root, i.e. it is a type frequency variable. We suppose that a large number of compounds with a particular root in the left-hand side of a compound (regardless of their token frequency) may impose the shape of that root on the right-hand side of a compound (again, following the same rationale as for

the previous hypothesis). The literature shows that this "family size" aspect may have an effect on (processing of) morphological structure (e.g. de Jong et al. (2000)). We hypothesize that both larger family size as well as higher family frequency of a root as the *left*-hand member of a compound (i.e. in its non-rendaku form) correlate with avoidance of rendaku ("hatred"). Conversely, we expect that both higher family size as well as higher family frequency of a root as the *right*-hand member of a compound (in its rendaku form) correlate with likelihood of rendaku occurring ("love"). Note, finally, that family size and family frequency are logically independent types of frequency: both, either, or none could be found to play a role in rendaku.

So, three types of frequency are hypothesized to interact with the propensity for rendaku in the sense that higher frequency scores are expected to block or reduce rendaku. In (8), we summarize the relevant hypotheses related to rendaku "hatred", together with the types of frequency on which they are based.

(8) *Hypothesis 1*

The frequency of a root in isolation is negatively correlated with its propensity for rendaku.

Hypothesis 2

The family size of the root as the left-hand member in compounds is negatively correlated with its propensity for rendaku.

Hypothesis 3

The family frequency of a root as the left-hand member in compounds is negatively correlated with its propensity for rendaku.

On the other hand, two types of frequency are hypothesized to interact with rendaku in the sense that higher frequencies are expected to enhance rendaku ("love").

(9) *Hypothesis 4*

The family size of the root as the right-hand member in compounds is positively correlated with its propensity for rendaku.

Hypothesis 5

The family frequency of the root as the right-hand member in compounds is positively correlated with its propensity for rendaku.

Apart from word frequency, we also took into account three other factors which might play a role but about which we had no strong *a priori* expectations: mora frequency, the number of syllables at the left-hand side of the compound, the occurrence of another voiced obstruent in the left-hand side member of the compound. First, we suspect that the frequency of the different initial moras involved in rendaku, e.g. [fu] and [bu] in *huro* and *buro*, respectively, have an influence on the propensity of rendaku. Perhaps some moras are favoured (i.e., more

frequent) than others in Japanese and therefore preferred (cf. Pierrehumbert (2006: 526)). With respect to this factor, other considerations could also come into play, e.g. the preference for a voiced allophone in intervocalic position or the general preference for voiceless obstruents over voiced ones. The second factor was the number of syllables on the left-hand side of the compound: is rendaku more likely to occur after shorter or longer roots? Thirdly, we included in the database whether another voiced obstruent was contained in the left-hand member of the compound (cf. extended Lyman's Law, see above). These three factors were treated as random variables, we will return to them in §4.4. In the next section, we will describe the methodology which we used to test the hypotheses in (8) and (9).

4.3 Methodology

To investigate the relevance of frequency for the lexical variation in rendaku, we selected a number of roots based on the appendices in Rosen (2001) (§4.3.1). Subsequently, we constructed a database of around 2,700 compounds in which these roots appear as either a left-hand member or a right-hand member (§4.3.2). Finally, we added frequency information of the roots in isolation and their frequency and family size as a left-hand side member and righ-hand side member of the compounds to each compound (§4.3.3).

4.3.1 Selection of the roots

Appendices D-F in Rosen (2001) form the starting-point of our investigation. These appendices contain 113 "*lovers*" (which (almost) always undergo rendaku), 19 "*haters*" (which (almost) never undergo rendaku – Rosen refers to these as the "never" category), and 4 "*doubters*" (which show variation, and which Rosen refers to as the "hater" category). Since our investigation focuses on the latter two categories, we intended to include all compounds with the roots that are marked as "haters" or "doubters" in Rosen (2001). However, we were forced to leave out two of the 19 "haters" (viz. *kera* 'mole cricket' and *kasa* 'shade'), because there were no compounds with these roots in the database created by Ogawa et al. (2005), which was consulted in the second phase of the construction of our database (see the next section). The final sets of 17 "haters" and four "doubters" are given in (10a) and (10b), respectively:

(10) *a.* *Rendaku "haters" (roots that almost never undergo rendaku) in the database*

1. sio	'tide'	10. himo	'string'
2. kata	'shoulder'	11. kami	'above'
3. hime	'princess'	12. take	'mushroom'
4. take	'measure'	13. siro	materials'
5. kuso	'dung'	14. kasu	'dregs'
6. suso	'cliff'	15. koi	'love'
7. saki	'tip'	16. tuti	'earth'
8. kase	'shackles'	17. tuyu	'dew'
9. tami	'people'		

b. *Rendaku doubters (roots that show rendaku variation)*

1. hara	'field'	3. kusa	'grass'
2. kawa	'skin'	4. kuse	'habit'

The large number of rendaku "lovers" in Rosen (2001) made it necessary to make a selection of "lovers" that satisfied a number of conditions relevant for our purposes. A number of these conditions are connected to the segmental and prosodic composition of these roots. In order to control for any influence of the segmental material of the first mora of these roots (see §4.2), we tried to include the same number (viz. 2) of roots for each initial consonant-vowel combination. Given the fact that there are four consonants that may undergo rendaku (/t s k h/), and a total of five vowels in Japanese, our ideal set of "lovers" consisted of $2 \times 4 \times 5 = 40$ roots. Unfortunately, some CV-combinations are not included in Rosen (2001) appendix (e.g. roots starting with /ti/), or represented by only one item (e.g. roots starting with /hi/). Therefore, seven forms in our "lovers" database were randomly selected from the remaining "lovers", without paying attention to the initial consonant. This resulted in the following set of "lovers":

Table 4.1. Forty rendaku "lovers" in the database, systematically varied for the consonant and the vowel of the first mora.

	/t/	/s/	/k/	/h/
/i/	––	sima 'stripe' siri 'buttocks' siru 'soup'	kimo 'kidney' kiri 'mist'	hire 'fin'
/e/	tera 'temple'	semi 'locust'	keta 'beam'	hera 'spatula' heri 'rim'
/a/	tama 'ball' tana 'shelf' tane 'seed'	sake 'alc. drink' sato 'village'	kai 'shell' kami 'paper'	hai 'ash' hako 'box' hato 'pigeon' hana 'flower' hana 'nose'
/o/	toki 'time' tori 'bird'	soko 'bottom' sono 'garden' sora 'sky'	koke 'moss' kosi 'hips'	hone 'bone' hosi 'star'
/u/	tuna 'rope' tuti 'hammer' tutu 'pipe'	sumi 'ink'	kuma 'bear' kutu 'shoes'	hue 'flute' huta 'lid'

To check whether the frequency of the initial mora might have an effect on rendaku variation, we included corpus-based frequency information for every mora, provided by Tamaoka & Makioka (2004).

In all, our database therefore consisted of 61 roots in total (17 "haters", 4 "doubters", and 40 "lovers"). Having established the set of roots which we investigated, let us now turn to the compounds in which these roots appear.

4.3.2 Selection of the compounds

As a second step, we collected as many compounds as possible in which the roots in our database actually occur and that were relevant for our investigation. For this, we used the Ogawa et al. (2005) database, which contains all 78,426 two-kanji words extracted from the fourth edition of the Kōjien dictionary (Kōjien 1991). All characters in these compounds belong to the set of 2,965 standard characters (Japanese Industrial Standard Level 1). By restricting the investigation to two-kanji roots, we obtained only compounds of two members, thus avoiding any effect of branching in longer compounds (see §2). In order to control for the other factors which might influence variation (see again §2), we excluded all compounds with a Sino-Japanese morpheme as the left member. Similarly, all compounds with left-hand members containing a long vowel were excluded. Compounds ending in the moraic nasal /N/

were discarded as well, since post-nasal voicing makes it impossible to determine whether voicing of an initial consonant on the right-hand side is due to rendaku or not.

This resulted in a data set of 2,702 compounds in which the 61 roots appear. In the next section, we explain how we collected frequency information on these roots and compounds.

4.3.3 Frequency

We included the following compound-dependent frequency types in the database (cf. the hypotheses in (8) and (9) above).

(11) a. Token frequency of the roots under investigation in isolation.

b. Token frequency of each compound.

c. Family size left: The number of different compounds in which a particular root appears as the left-hand member.

d. Family frequency left: The sum of all token frequencies of all compounds in which a particular root appears as the left-hand member.

e. Family size right: The number of different compounds in which a particular root appears as the right-hand member.

f. Family frequency right: The sum of all token frequencies of all compounds in which a particular root appears as the right-hand member.

All token frequencies of roots and compounds were computed on the basis of the Chunagon database of the Balanced Corpus of Contemporary Written Japanese (BCCWJ 2011). We used the online version of the database, which is based on a corpus of written Japanese containing texts written between 1971 and 2008, taken from all kinds of genres: newspapers, magazines, textbooks, poetry, PR brochures, legal texts, recordings of meetings of the Japanese Diet, and the internet (such as blogs). This online version contains ten million words. All possible different spellings (if applicable) were included in the frequency count. The frequencies of compounds in the database ranged from 0 (e.g. *imogai* 'cone shell') to 7,147 (*tegami* 'letter'). It should be noted that, since the database is ultimately based on a general dictionary, it contains many infrequent words.[2]

All frequency counts were log-transformed, since frequency is processed in memory in logarithmic values rather than absolute values (Shapiro 1969). The different log-transformed frequencies vary as follows:

[2] For many words in the corpus, the token frequency is zero. In the log transformation, the value was also set on zero.

(12)

Root in isolation	< 5.2
Compounds	< 3.9
Family Frequency Right	< 3.9
Family Frequency Left	< 3.9
Family Size Right	< 2.1
Family Size Left	< 1.8

These figures mean that the frequency range of roots in isolation was largest and the frequency range for Family Size Left was smallest.

4.4 Results

We performed a logistic regression test (R Development Core Team (2009), Baayen (2008)) with the different frequency factors as variables (i.e. root frequency in isolation, family frequency and family size of the root as a left-hand side member and as a right-hand side member of the compound). We also checked the random variables of initial mora and its frequency, the number of syllables on the left-hand side of the compound, and the presence of another voiced segment in the left-hand member of the compound (see §4.2). The frequency of the root in isolation was positively correlated with the propensity of the root for undergoing rendaku ($z = 6.499$, $p < 0.001$). Similarly, the family size of the root as a left-hand side member was positively correlated with the likelihood of the root for undergoing rendaku ($z = 4.884$, $p < 0.001$). On the other hand, the family frequency of the root as a left-hand side member was *negatively* correlated with the probability of the root for undergoing rendaku ($z = -10.17$, $p < 0.001$). Mora frequency was negatively correlated with rendaku, that is, the higher the mora frequency, the less likely it is to undergo rendaku ($z = -4.549$, $p < 0.001$). Finally, the number of syllables on the left-hand side of the compound turned out to be significant ($z = -6.096$, $p < 0.001$). The results are summarized in Table 4.2.

Table 4.2. Results of rendaku: the optimal logistic regression results, showing the estimates, S.E, z-value, and p-value for the log frequency of the root in isolation, the log family size and the family frequency of the root as a left-hand side member of the compound, the frequency of the initial mora of the root, and the number of syllables at the left-hand side of the compound.

	Est.	S.E.	z-value	p-value
(Intercept)	−2.609	0.456	−2.029	0.043
Log Frequency Isolation	0.816	0.128	6.499	<0.001*
Log Family Size Left	0.909	0.184	4.884	<0.001*
Family Frequency Left*1000	−0.481	0.047	−10.17	<0.001*
Frequency Mora*1000	−0.004	−0.001	−4.549	<0.001*
Number of syllables	0.179	0.105	−6.096	<0.001*

Let us consider some concrete examples. Table 4.2 shows that roots with a low frequency in isolation, such as *siro* 'materials' (log frequency in isolation 0.8), are unlikely to undergo rendaku and roots with a higher frequency in isolation, such as *toki* 'time' (log frequency in isolation 5.2) are more likely to undergo rendaku. For example, in our database, *siro* undergoes rendaku in 2 out of 55 cases (3.6%) and *toki* undergoes rendaku in 51 out of 58 cases (75.9%). Further, although rendaku is concerned with nouns on the right-hand side of compounds, the results show that we should also take into account the occurrence of such roots when they appear on the *left-hand* side of compounds if we wish to understand the behaviour of "lovers", "haters", and "doubters" in rendaku. Family size of the root as a left-hand side member of roots positively correlates with rendaku. When we compare *kuse* 'habit' with *tama* 'ball', both have a log frequency in isolation of 3.6, we find that *kuse* (log family size left 0.5), undergoes rendaku in 7 out of 19 cases (36.8%), whereas *tama* (log family size left 1.8) undergoes rendaku in 40 out of 52 cases (76.9%).

We hypothesized that the frequency of the root on the left side of the compound has a negative effect on rendaku. This hypothesis is born out. In an Exemplar Theory approach, this makes sense, because this non-rendaku allomorph (with a voiceless consonant) will have a mental representation that is relatively strong compared to the rendaku form (with a voiced consonant). However, the other hypotheses initially spelt out in (8) and (9) are not confirmed. First, we did not find any effect of the frequency of the root as a right-hand member of the compound (either in terms of family size or of family frequency). Moreover, our hypotheses about the negative effect of the root in isolation and family size at the left side of the compound, which would lead to less rendaku in right-hand position, are even disconfirmed. Unexpectedly, higher root frequency leads to more rendaku and likewise larger family size of the root as a left-hand side member also leads to more rendaku. Note, besides, that the frequencies in our database are relatively low (log frequency <5.2 out of a corpus of 10 million

words, see (12)), which indicates that the variation in rendaku occurs exactly in very infrequent words. In the following section, we will try to interprete and model the results in EPOT.

4.5 Rendaku analysis in EPOT

We saw that rendaku voices the first consonant of the second part of a compound—which can be characterized as a phonological rule. This process can be captured either by a (number of) (morpho-)phonological rules or as a result of constraint interaction (see, among others, Itô & Mester (2003)). That is, the process is partly subject to purely phonological conditions, some of which are near-categorical (e.g. Lyman's Law, see (2)). The variation to which rendaku is subject should, however, not be ignored. In order to explain this kind of lexical variation, i.e. whether Japanese roots behave as "rendaku-lovers" or "rendaku-haters", we found that frequency of use and morphological family size are relevant. This holds especially true for cases in which rendaku fails to apply. In other words, in order to provide a full account for rendaku we need to make reference to phonological rules, i.e. grammar, as well as to frequency. Thus, *both* the frequency of particular items *and* phonological rules (or constraints) play a role in the application (or non-application) of this process.

We should emphasize that variation occurs only in a small subset of the Japanese lexicon— it affects mostly roots that occur in infrequent compounds. In frequency studies, it is well known that high-frequency word often behave in an idiosyncratic way, due to an extremely strong mental representation, which prevents (analogical) change throughout time (Bybee 2002). However, our results show that low-frequency words may also behave differently. Noticeably, this frequency effect does not resemble the two well-known frequency effects mentioned in §1.4: rendaku does not involve reduction of the roots, so it cannot show a Type II frequency effect (reduction, which affects HF words first); nor can the results be attributed to a Type I frequency effect of analogical change which affect LF words first (see the introduction in chapter 1). The frequency effect we found in rendaku differs and we call this Type III frequency effect (opaque structure). In the following chapter, we will consider another case in which low-frequency words are exceptional with regard to a particular pattern of change: lowering of long <ä> in German.

Since we observed a phonological rule as well as usage-based effects, rendaku forms a good test-case for EPOT. The tasks of EPOT are (1) to account for the frequency effects, (2) to decide on the input, and (3) to account for the grammatical rule. Let us first consider the frequency effects, which are modelled in the lexical component of EPOT. Since frequency effects occur especially in LF words (and words with higher frequency undergo the rendaku rule), we tentatively suggest that the morphological status of compounds with roots that are extremely infrequent may be unclear for some speakers. In order to recognize morphological patterns and regularities, a reasonable number of similar words must be stored in the lexicon. Some of the words in our database might not be recognized as compounds and therefore it might be unclear whether rendaku should be applied or not. Suppose a compound with

extremely low frequency and a small family size at the left side of the compound is perceived; the listener may not realize that this is a compound and will store it as a monomorphemic word. Without the morpheme boundary being recognized, (almost) no connections with the morphemes in other compounds in the lexicon will be made. For instance, a typical "hater" like *tumisiro* 'atonement' (low log frequency in isolation 0.8 and low log family size 0.6) may not be treated as a compound by the language system, hence it does not undergo the rule of rendaku. Also, the log family frequency left is moderate (2.7), which will lead to lexical competition between the different exemplars and give the voiceless variant a relatively strong representation. A typical case of a "doubter" is *kusa* 'grass' (which has a moderate frequency of the root in isolation 3.5, a moderate family frequency left frequency 2.1, as well as a moderate family size left frequency 2.7). Depending on the outcome of the lexical competition between the different forms, sometimes, the root in the "doubter" will be recognized as such and sometimes it will not be recognized as such. Therefore such a compounds vacillates. Usually, however, roots occur frequently enough to be recognized as such in compounds, and so they will appear as "lovers".

So, we assume that the root frequencies correlate with the probability that a root can be recognized as a separate morpheme. If enough exemplars in the lexicon of a root in isolation or of the root as a left hand member of the compound ocuur, the compound will be recognized as a morphological complex word. If not enough of such exemplars are present in the lexicon, the compound will be stored as an exemplar of a monomorphemic root, without connections to exemplars of roots. In other words, as for the structure of the lexicon, we suppose that whether a compound undergoes rendaku or not crucially depends on the strength of the connections between exemplars. HF roots have strong connections and will be easily recognized as a separate morpheme within a compound. LF roots, on the other hand, have looser connections (some connections on the basis of similar forms will always exist). The exemplar storage is illustrated in Figure 4.1a for the LF root *siro* substitution' and in Figure 4.1b for the root *hue* 'flute'. The illustrations are simplified for reasons of clearity. Numerous other connections exist, such as the connections between all words in the figures and the connections with roots on the left-hand side of the compounds.

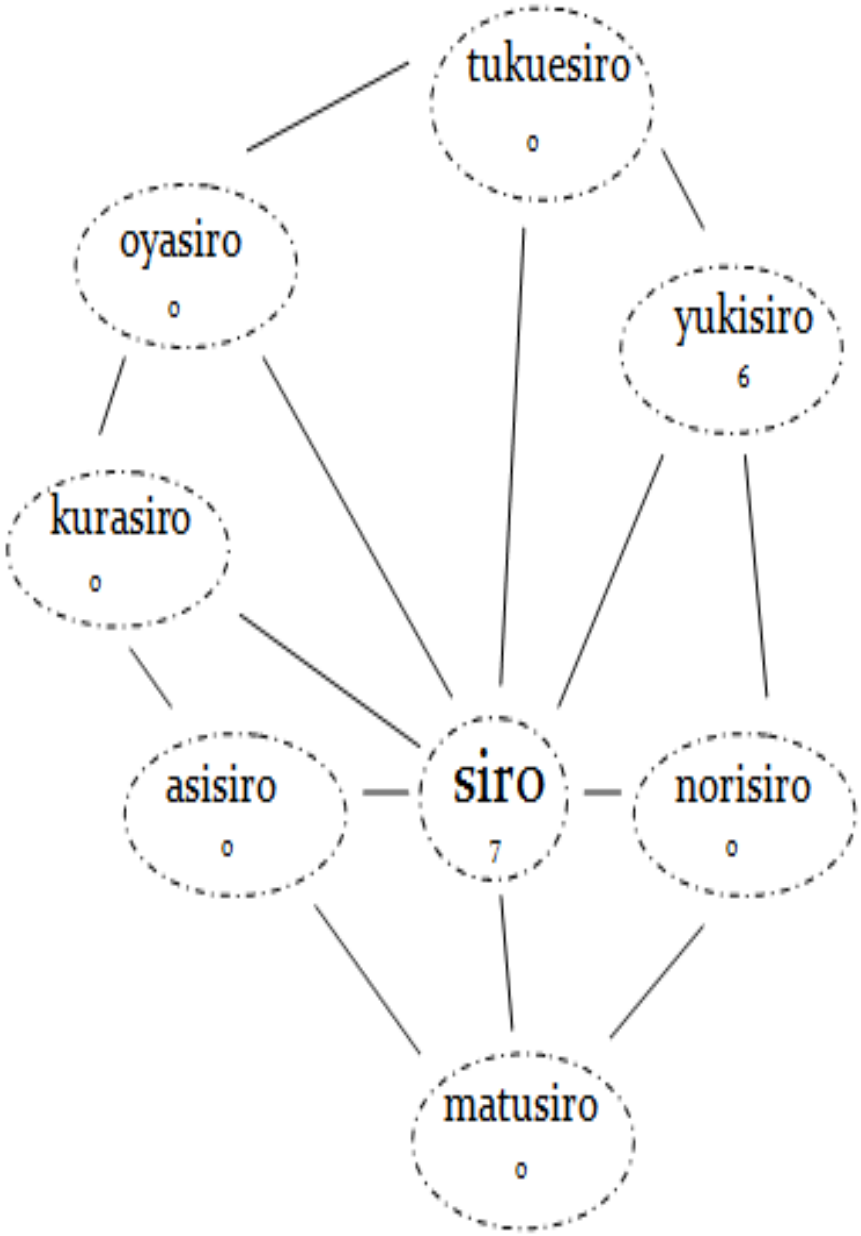

Figure 4.1a. Exemplars and connections of the LF root 'siro' and compounds with 'siro'. The relative strength of the exemplar categories is indicated by the size of the letters and the relative strength of the exemplar connections is indicated by the width of the lines. The frequency of the words is given in the exemplar clouds. The connections are all very loose.

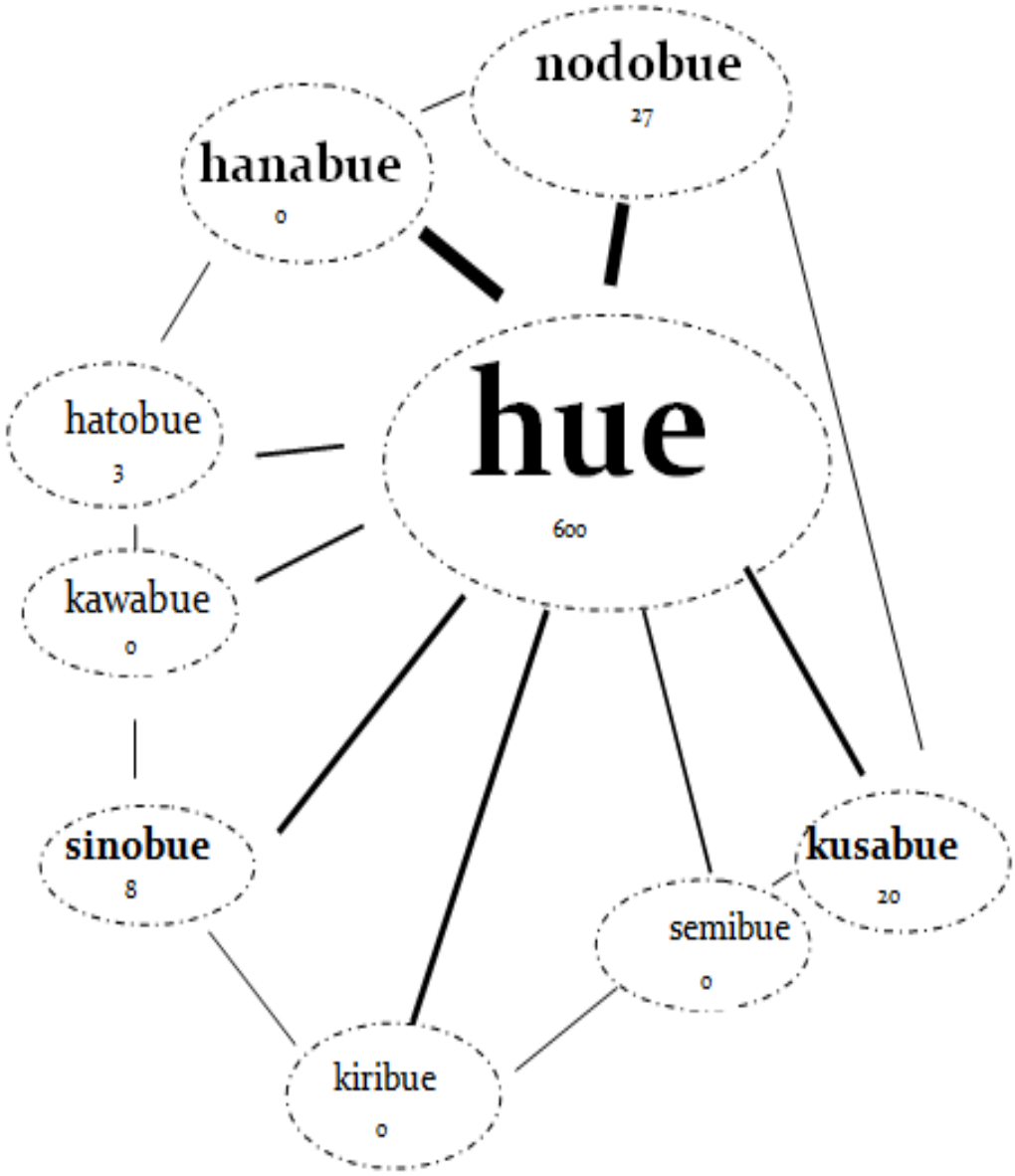

Figure 4.1b. Exemplars and connections of the HF root hue and compounds with hue. The relative strength of the exemplar categories is indicated by the size of the letters and the relative strength of the exemplar connections is indicated by the width of the lines. The frequency of the words is given in the exemplar clouds. The connections are relatively strong.

Given the difference in representation for LF and HF words in the lexicon, the next question is how EPOT selects an input for the grammar. As explained above, the morphological structure of the word crucially depends on the interplay of different types of frequency. The input is determined by the activation of a particular root or variant of that root. The extremely LF compounds are supposed to be stored more or less separately from the root and other compounds containing that root, and will thus appear as a monomorphemic unit in the input. Words that have higher frequencies are supposed to be recognized as compounds and their input will thus consist of two roots. The right-hand side will be formed by the *prototype* of the root. Roots in isolation are the most frequent, thus these variants (the non-rendaku form) will have the highest activation. Therefore, the non-rendaku form will be selected in the input. Thus, the extremely infrequent word *tumisiro* 'atonement' will appear in the input as $[\text{tumisiro}]_N$ and the more frequent *sinobue* will appear in the input as $[[\text{sino}]_N[\text{hue}]_N]_N$ 'bamboo flute'.

As for the OT account, I follow Ito & Mester (1986, 1998) and Fukuzawa & Kitahara (2001). Ito & Mester (1986, 1998) proposed that rendaku can be best understood as a linking morpheme that bears the feature [voice] (recall that, historically, the linking morpheme was the possessive particle *no*, which got reduced to a single feature [voiced]).

(13)

Stem 1		Stem 2
sino	ρ	hue
	|	
	[voice]	

The relevant constraints are REALIZE-MORPHEME and IDENT[voice].

(14) REALIZE-MORPHEME

A morpheme in the input should be realized in the output.

(15) IDENT[voice]

Assign a violation mark to any output segment that does not correspond in voicing with an input segment.

Further, Lyman's Law is analysed as an OCP rule (Obligatory Contour Principle) on [voice] that blocks two voiceless consonants in the compound.

(16) OCP[voice]

Assign a violation mark to a morpheme that contains two voiced consonants.

Tableau (17) shows the evaluation of *sinohue* and *tumisiro*.

(17) *Tableaux for sinohue, tumisiro, and kamikaze.*

$[[sino]_N$ ρ $[hue]_N]_N$	OCP[voice]	REALIZE-MORPHEME	IDENT[voice]
sinohue		*!	
☞ sinobue			*

$[tumisiro]_N$	OCP[voice]	REALIZE-MORPHEME	IDENT[voice]
☞ tumisiro			
tumiziro			*!

$[[kami]_N$ ρ $[kaze]_N]_N$	OCP[voice]	REALIZE-MORPHEME	IDENT[voice]
☞ kamikaze		*	
kamigaze	*!		*

The tableaux show the effect of the grammar on inputs with different morphological structure which provides an output with rendaku in case of a morphologically complex input and an output without rendaku in case of a monomorphemic input.

4.6 Conclusion

In this chapter, I presented an example how EPOT accounts for the modelling of rendaku. Rendaku is a phonological rule in Japanese compounding, but we found that LF words, due to their opaque structure, do not always undergo the rule. The lexicon is modelled on the basis of the frequency effects that we observed. I provided (simplified) illustrations of the lexical representation of a root with relatively strong lexical connections and a lexical representation of a root with relatively weak connections. Lexical strength depends on the frequency of the root in isolation and the occurrence of the root as a left-hand side member of compounds. Compounds with stronger connections are supposed to be morphologically analysed as complex words, i.e. the root is recognized as such. On the other hand, compounds with

weaker connections are supposed to be stored as monomorphemic units, i.e. the root is not recognized as such. Subsequently, I proposed that EPOT selects the input on the basis of the difference in lexical organization (due to different frequency values) of these compounds. Either a monomorphemic word is the input, or the compound with the root as it appears in isolation is the input. The—invariable—EPOT grammar provides outputs that vary only on the basis of the morphological structure of the input.